Aidetic Data Engineer - Assignment Pyspark 2023

Assignment Description

In this assignment, I will demonstrate my ability to work with Pyspark

Assignment Instructions:

● Set up your development environment with the necessary libraries.

● Access the provided dataset from here and save it into your local directory.

● Setup a local mysql database.

● Create a new table named neic_earthquakes.

● Run a python script to read the given data and push the data into the neic_earthquakes table.

● Read the data from the table into a PySpark DataFrame and answer the following questions:

○ How does the Day of a Week affect the number of earthquakes?

○ What is the relation between Day of the month and Number of earthquakes that happened in a year?

○ What does the average frequency of earthquakes in a month from the year 1965 to 2016 tell us?

○ What is the relation between Year and Number of earthquakes that happened in that year?

○ How has the earthquake magnitude on average been varied over the years?

○ How does year impact the standard deviation of the earthquakes?

○ Does geographic location have anything to do with earthquakes?

○ Where do earthquakes occur very frequently?

○ What is the relation between Magnitude, Magnitude Type , Status and Root Mean Square of the earthquakes?

1. Install Libraries

In [1]:
```
pip install pandas pymysql findspark
```

```
Requirement already satisfied: pandas in c:\users\nithi\anaconda3\conda\lib\site-packages (2.0.3)
Requirement already satisfied: pymysql in c:\users\nithi\anaconda3\conda\lib\site-packages (1.1.0)
Requirement already satisfied: findspark in c:\users\nithi\anaconda3\conda\lib\site-packages (2.0.1)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\users\nithi\anaconda3\conda\lib\site-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in c:\users\nithi\anaconda3\conda\lib\site-packages (from pandas) (2022.7)
Requirement already satisfied: tzdata>=2022.1 in c:\users\nithi\anaconda3\conda\lib\site-packages (from pandas) (2023.3)
Requirement already satisfied: numpy>=1.21.0 in c:\users\nithi\anaconda3\conda\lib\site-packages (from pandas) (1.24.3)
Requirement already satisfied: six>=1.5 in c:\users\nithi\anaconda3\conda\lib\site-packages (from python-dateutil>=2.8.2->pandas) (1.16.0)
Note: you may need to restart the kernel to use updated packages.
[notice] A new release of pip is available: 23.2.1 -> 23.3.1
[notice] To update, run: python.exe -m pip install --upgrade pip
```

Python Script to Load Data into MySQL

Lets create a Python script, load_data.py, to read the provided data and push it into the MySQL database. Use the pandas library for reading data and pymysql for database connectivity.

In [2]:
```
import pandas as pd
import pymysql
from sqlalchemy import create_engine

# Read data into a pandas DataFrame
data = pd.read_csv(r'E:\neic_earthquake\database.csv')

# Connect to MySQL
#engine = create_engine('mysql+pymysql://root:"Nithin@2001"@localhost/MySQL')
engine = create_engine("mysql+pymysql://root:Nithin2001@localhost/MySQL")

# Push data to MySQL
data.to_sql('neic_earthquakes', con=engine, if_exists='replace', index=False)
```

Out[2]:
```
23412
```

PySpark Analysis:

Create another Python script to analyze the data using PySpark. Lets call it analyze_data.py.

In [3]:
```
from pyspark.sql import SparkSession

# Initialize Spark session
spark = SparkSession.builder.appName("EarthquakeAnalysis").getOrCreate()
```

In [4]:
```
spark
```

Out[4]:

SparkSession - in-memory

SparkContext

Spark UI

Version	`v3.5.0`
Master	`local[*]`
AppName	`EarthquakeAnalysis`

In [5]:
```
df = spark.read.csv("E:\\neic_earthquake\\database.csv",header=True,inferSchema=True)
```

In [6]:
```
df.show()
```

```
+---------+------------------+--------+--------+---------+-----+----------+-----
----------------+---------+-------------+--------------+-------------------------
+------------+------------------+---------------+---------------+---------------
----+---------+--------------+---------------+---------+
|      Date|              Time|Latitude|Longitude|      Type|Depth|Depth Error|Depth
Seismic Stations|Magnitude|Magnitude Type|Magnitude Error|Magnitude Seismic Stations|
Azimuthal Gap|Horizontal Distance|Horizontal Error|Root Mean Square|
ID|   Source|Location Source|Magnitude Source|   Status|
+---------+------------------+--------+--------+---------+-----+----------+-----
----------------+---------+-------------+--------------+-------------------------
+------------+------------------+---------------+---------------+---------------
----+---------+--------------+---------------+---------+
|01/02/1965|2023-11-21 13:44:18|  19.246|  145.616|Earthquake|131.6|       NULL|
NULL|      6.0|            MW|           NULL|                      NULL|         NUL
L|               NULL|            NULL|            NULL|        ISCGEM860706|   ISCGE
M|         ISCGEM|          ISCGEM|Automatic|
|01/04/1965|2023-11-21 11:29:49|   1.863|  127.352|Earthquake| 80.0|       NULL|
NULL|      5.8|            MW|           NULL|                      NULL|         NUL
L|               NULL|            NULL|            NULL|        ISCGEM860737|   ISCGE
M|         ISCGEM|          ISCGEM|Automatic|
|01/05/1965|2023-11-21 18:05:58| -20.579| -173.972|Earthquake| 20.0|       NULL|
NULL|      6.2|            MW|           NULL|                      NULL|         NUL
L|               NULL|            NULL|            NULL|        ISCGEM860762|   ISCGE
M|         ISCGEM|          ISCGEM|Automatic|
|01/08/1965|2023-11-21 18:49:43| -59.076|  -23.557|Earthquake| 15.0|       NULL|
NULL|      5.8|            MW|           NULL|                      NULL|         NUL
L|               NULL|            NULL|            NULL|        ISCGEM860856|   ISCGE
M|         ISCGEM|          ISCGEM|Automatic|
|01/09/1965|2023-11-21 13:32:50|  11.938|  126.427|Earthquake| 15.0|       NULL|
NULL|      5.8|            MW|           NULL|                      NULL|         NUL
L|               NULL|            NULL|            NULL|        ISCGEM860890|   ISCGE
M|         ISCGEM|          ISCGEM|Automatic|
|01/10/1965|2023-11-21 13:36:32| -13.405|  166.629|Earthquake| 35.0|       NULL|
NULL|      6.7|            MW|           NULL|                      NULL|         NUL
L|               NULL|            NULL|            NULL|        ISCGEM860922|   ISCGE
M|         ISCGEM|          ISCGEM|Automatic|
|01/12/1965|2023-11-21 13:32:25|  27.357|   87.867|Earthquake| 20.0|       NULL|
NULL|      5.9|            MW|           NULL|                      NULL|         NUL
L|               NULL|            NULL|            NULL|        ISCGEM861007|   ISCGE
M|         ISCGEM|          ISCGEM|Automatic|
|01/15/1965|2023-11-21 23:17:42| -13.309|  166.212|Earthquake| 35.0|       NULL|
NULL|      6.0|            MW|           NULL|                      NULL|         NUL
L|               NULL|            NULL|            NULL|        ISCGEM861111|   ISCGE
M|         ISCGEM|          ISCGEM|Automatic|
|01/16/1965|2023-11-21 11:32:37| -56.452|  -27.043|Earthquake| 95.0|       NULL|
NULL|      6.0|            MW|           NULL|                      NULL|         NUL
L|               NULL|            NULL|            NULL|     ISCGEMSUP861125|ISCGEMSU
P|         ISCGEM|          ISCGEM|Automatic|
|01/17/1965|2023-11-21 10:43:17| -24.563|  178.487|Earthquake|565.0|       NULL|
NULL|      5.8|            MW|           NULL|                      NULL|         NUL
L|               NULL|            NULL|            NULL|        ISCGEM861148|   ISCGE
M|         ISCGEM|          ISCGEM|Automatic|
|01/17/1965|2023-11-21 20:57:41|  -6.807|  108.988|Earthquake|227.9|       NULL|
NULL|      5.9|            MW|           NULL|                      NULL|         NUL
L|               NULL|            NULL|            NULL|        ISCGEM861155|   ISCGE
M|         ISCGEM|          ISCGEM|Automatic|
|01/24/1965|2023-11-21 00:11:17|  -2.608|  125.952|Earthquake| 20.0|       NULL|
NULL|      8.2|            MW|           NULL|                      NULL|         NUL
L|               NULL|            NULL|            NULL|        ISCGEM861299|   ISCGE
M|         ISCGEM|          ISCGEM|Automatic|
```

```
|01/29/1965|2023-11-21 09:35:30|  54.636|  161.703|Earthquake| 55.0|       NULL|
NULL|      5.5|            MW|           NULL|                      NULL|         NUL
L|               NULL|            NULL|            NULL|        ISCGEM861461|   ISCGE
M|         ISCGEM|          ISCGEM|Automatic|
|02/01/1965|2023-11-21 05:27:06| -18.697| -177.864|Earthquake|482.9|       NULL|
NULL|      5.6|            MW|           NULL|                      NULL|         NUL
L|               NULL|            NULL|            NULL|        ISCGEM859136|   ISCGE
M|         ISCGEM|          ISCGEM|Automatic|
|02/02/1965|2023-11-21 15:56:51|  37.523|   73.251|Earthquake| 15.0|       NULL|
NULL|      6.0|            MW|           NULL|                      NULL|         NUL
L|               NULL|            NULL|            NULL|        ISCGEM859164|   ISCGE
M|         ISCGEM|          ISCGEM|Automatic|
|02/04/1965|2023-11-21 03:25:00|  -51.84|  139.741|Earthquake| 10.0|       NULL|
NULL|      6.1|            MW|           NULL|                      NULL|         NUL
L|               NULL|            NULL|            NULL|        ISCGEM859200|   ISCGE
M|         ISCGEM|          ISCGEM|Automatic|
|02/04/1965|2023-11-21 05:01:22|  51.251|  178.715|Earthquake| 30.3|       NULL|
NULL|      8.7|            MW|           NULL|                      NULL|         NUL
L|               NULL|            NULL|            NULL|OFFICIAL196502040...| OFFICIA
L|         ISCGEM|        OFFICIAL|Automatic|
|02/04/1965|2023-11-21 06:04:59|  51.639|  175.055|Earthquake| 30.0|       NULL|
NULL|      6.0|            MW|           NULL|                      NULL|         NUL
L|               NULL|            NULL|            NULL|     ISCGEMSUP859215|ISCGEMSU
P|         ISCGEM|          ISCGEM|Automatic|
|02/04/1965|2023-11-21 06:37:06|  52.528|  172.007|Earthquake| 25.0|       NULL|
NULL|      5.7|            MW|           NULL|                      NULL|         NUL
L|               NULL|            NULL|            NULL|        ISCGEM859221|   ISCGE
M|         ISCGEM|          ISCGEM|Automatic|
|02/04/1965|2023-11-21 06:39:32|  51.626|  175.746|Earthquake| 25.0|       NULL|
NULL|      5.8|            MW|           NULL|                      NULL|         NUL
L|               NULL|            NULL|            NULL|        ISCGEM859222|   ISCGE
M|         ISCGEM|          ISCGEM|Automatic|
+----------+-------------------+--------+---------+----------+-----+-----------+-----
----------------+---------+-------------+---------------+-------------------------
+-------------+-------------------+----------------+----------------+----------------
----+---------+---------------+----------------+---------+
only showing top 20 rows
```

In [7]:
```
df.head(2)
```

Out[7]:
```
[Row(Date='01/02/1965', Time=datetime.datetime(2023, 11, 21, 13, 44, 18), Latitude=1
9.246, Longitude=145.616, Type='Earthquake', Depth=131.6, Depth Error=None, Depth Sei
smic Stations=None, Magnitude=6.0, Magnitude Type='MW', Magnitude Error=None, Magnitu
de Seismic Stations=None, Azimuthal Gap=None, Horizontal Distance=None, Horizontal Er
ror=None, Root Mean Square=None, ID='ISCGEM860706', Source='ISCGEM', Location Source
='ISCGEM', Magnitude Source='ISCGEM', Status='Automatic'),
 Row(Date='01/04/1965', Time=datetime.datetime(2023, 11, 21, 11, 29, 49), Latitude=1.
863, Longitude=127.352, Type='Earthquake', Depth=80.0, Depth Error=None, Depth Seismi
c Stations=None, Magnitude=5.8, Magnitude Type='MW', Magnitude Error=None, Magnitude
Seismic Stations=None, Azimuthal Gap=None, Horizontal Distance=None, Horizontal Error
=None, Root Mean Square=None, ID='ISCGEM860737', Source='ISCGEM', Location Source='IS
CGEM', Magnitude Source='ISCGEM', Status='Automatic')]
```

In [8]:
```
df.printSchema()
```

```
root
 |-- Date: string (nullable = true)
 |-- Time: timestamp (nullable = true)
 |-- Latitude: double (nullable = true)
 |-- Longitude: double (nullable = true)
 |-- Type: string (nullable = true)
 |-- Depth: double (nullable = true)
 |-- Depth Error: double (nullable = true)
 |-- Depth Seismic Stations: integer (nullable = true)
 |-- Magnitude: double (nullable = true)
 |-- Magnitude Type: string (nullable = true)
 |-- Magnitude Error: double (nullable = true)
 |-- Magnitude Seismic Stations: integer (nullable = true)
 |-- Azimuthal Gap: double (nullable = true)
 |-- Horizontal Distance: double (nullable = true)
 |-- Horizontal Error: double (nullable = true)
 |-- Root Mean Square: double (nullable = true)
 |-- ID: string (nullable = true)
 |-- Source: string (nullable = true)
 |-- Location Source: string (nullable = true)
 |-- Magnitude Source: string (nullable = true)
 |-- Status: string (nullable = true)
```

In [9]:
```
df.columns
```

Out[9]:
```
['Date',
 'Time',
 'Latitude',
 'Longitude',
 'Type',
 'Depth',
 'Depth Error',
 'Depth Seismic Stations',
 'Magnitude',
 'Magnitude Type',
 'Magnitude Error',
 'Magnitude Seismic Stations',
 'Azimuthal Gap',
 'Horizontal Distance',
 'Horizontal Error',
 'Root Mean Square',
 'ID',
 'Source',
 'Location Source',
 'Magnitude Source',
 'Status']
```

In [10]:
```
from pyspark.sql.functions import dayofweek, dayofmonth, month, year, avg, stddev, cou
```

In [11]:
```
# How does the Day of a Week affect the number of earthquakes?
df_day_of_week = df.groupBy(dayofweek("Date").alias("day_of_week")).count().orderBy("c
```

In [12]:
```
# What is the relation between Day of the month and Number of earthquakes that happene
df_day_of_month = df.groupBy(year("Date").alias("year"), dayofmonth("Date").alias("day
```

In [13]:
```
# What does the average frequency of earthquakes in a month from the year 1965 to 2016
df_avg_frequency = df.filter((year("Date") >= 1965) & (year("Date") <= 2016)).groupBy(
```

In [14]:
```
# What is the relation between Year and Number of earthquakes that happened in that ye
df_yearly_count = df.groupBy(year("Date").alias("year")).agg(count("*").alias("earthqu
```

In [15]:
```
# How has the earthquake magnitude on average been varied over the years?
df_avg_magnitude_over_years = df.groupBy(year("Date").alias("year")).agg(avg("Magnitud
```

In [16]:
```
# How does year impact the standard deviation of the earthquakes?
df_stddev_over_years = df.groupBy(year("Date").alias("year")).agg(stddev("Magnitude").
```

In [17]:
```
# Does geographic location have anything to do with earthquakes?
# Note: You may need to adapt this based on your specific dataset columns related to g
df_geo_analysis = df.groupBy("Latitude", "Longitude").agg(count("*").alias("earthquake
```

In [18]:
```
# Where do earthquakes occur very frequently?
df_frequent_locations = df.groupBy("Latitude", "Longitude").agg(count("*").alias("eart
```

In [19]:
```
# What is the relation between Magnitude, Magnitude Type, Status, and Root Mean Square
df_magnitude_relation = df.groupBy("Magnitude", "Magnitude Type", "Status", "Root Mear
```

In [20]:
```
# Print or save the results as needed
df_day_of_week.show()
df_day_of_month.show()
df_avg_frequency.show()
df_yearly_count.show()
df_avg_magnitude_over_years.show()
df_stddev_over_years.show()
df_geo_analysis.show()
df_frequent_locations.show()
df_magnitude_relation.show()
```

```
+-----------+-----+
|day_of_week|count|
+-----------+-----+
|       NULL|23409|
|          1|    3|
+-----------+-----+

+----+------------+-----+
|year|day_of_month|count|
+----+------------+-----+
|NULL|        NULL|23409|
|1975|          23|    1|
|1985|          28|    1|
|2011|          13|    1|
+----+------------+-----+

+-----+-------------+
|month|avg_Magnitude|
+-----+-------------+
|    2|          5.6|
|    3|          5.8|
|    4|          5.6|
+-----+-------------+

+----+----------------+
|year|earthquake_count|
+----+----------------+
|NULL|           23409|
|1975|               1|
|1985|               1|
|2011|               1|
+----+----------------+

+----+-----------------+
|year|    avg_Magnitude|
+----+-----------------+
|NULL|5.882558417702829|
|1975|              5.6|
|1985|              5.6|
|2011|              5.8|
+----+-----------------+

+----+------------------+
|year|  Magnitude_stddev|
+----+------------------+
|NULL|0.4230843439717061|
|1975|              NULL|
|1985|              NULL|
|2011|              NULL|
+----+------------------+

+--------+---------+----------------+
|Latitude|Longitude|earthquake_count|
+--------+---------+----------------+
|    51.5|   -174.8|               4|
|  34.416|  -118.37|               3|
|   38.64|   142.75|               2|
|  51.752|    175.5|               1|
|  -0.007|  125.026|               1|
|  37.702|   13.033|               1|
```

```
|  51.633|  159.325|               1|
|  40.406|  143.767|               1|
| -18.499|  -63.472|               1|
|  43.323|  147.713|               1|
|   8.187|  126.286|               1|
|  42.282|  143.224|               1|
| -20.248|  168.745|               1|
|   7.478|  123.809|               1|
|  33.362|  140.827|               1|
|  -1.592|  136.656|               1|
|  36.469|   70.912|               1|
| -55.506|  -28.309|               1|
|  18.208|  119.125|               1|
| -57.416|   -25.67|               1|
+--------+---------+----------------+
only showing top 20 rows

+--------+---------+----------------+
|Latitude|Longitude|earthquake_count|
+--------+---------+----------------+
|    51.5|   -174.8|               4|
|  34.416|  -118.37|               3|
|   38.64|   142.75|               2|
|  51.752|    175.5|               1|
|  -0.007|  125.026|               1|
|  37.702|   13.033|               1|
|  51.633|  159.325|               1|
|  40.406|  143.767|               1|
| -18.499|  -63.472|               1|
|  43.323|  147.713|               1|
|   8.187|  126.286|               1|
|  42.282|  143.224|               1|
| -20.248|  168.745|               1|
|   7.478|  123.809|               1|
|  33.362|  140.827|               1|
|  -1.592|  136.656|               1|
|  36.469|   70.912|               1|
| -55.506|  -28.309|               1|
|  18.208|  119.125|               1|
| -57.416|   -25.67|               1|
+--------+---------+----------------+
only showing top 20 rows

+---------+--------------+---------+----------------+-----+
|Magnitude|Magnitude Type|   Status|Root Mean Square|count|
+---------+--------------+---------+----------------+-----+
|      5.5|            MB| Reviewed|            NULL|  665|
|      5.6|            MB| Reviewed|            NULL|  475|
|      5.7|            MW|Automatic|            NULL|  393|
|      5.6|            MW|Automatic|            NULL|  366|
|      5.8|            MW|Automatic|            NULL|  332|
|      5.7|            MB| Reviewed|            NULL|  302|
|      5.5|            MW| Reviewed|             1.0|  263|
|      5.5|            MW| Reviewed|             1.1|  257|
|      6.0|            MW|Automatic|            NULL|  247|
|      5.9|            MW|Automatic|            NULL|  243|
|      5.6|            MW| Reviewed|             1.1|  222|
|      5.8|            MB| Reviewed|            NULL|  218|
|      5.6|            MW| Reviewed|             1.0|  213|
|      6.1|            MW|Automatic|            NULL|  178|
```

```
|      5.7|           MW| Reviewed|             1.0|  171|
|      5.7|           MW| Reviewed|             1.1|  167|
|      5.5|           MW| Reviewed|             1.2|  156|
|      5.5|           MW| Reviewed|             0.9|  152|
|      5.9|           MB| Reviewed|            NULL|  150|
|      6.2|           MW|Automatic|            NULL|  148|
+---------+-------------+---------+----------------+-----+
only showing top 20 rows
```

In [22]:
```
# Stop Spark session
spark.stop()
```